

英汉 语言测试 词典

An
English-Chinese
Dictionary
of Language
Testing

An
English-Chinese
Dictionary
of Language
Testing

An
English-Chinese
Dictionary
of Language
Testing

主编：王振亚



北京语言大学出版社
BEIJING LANGUAGE AND CULTURE
UNIVERSITY PRESS

ECDLT

英汉 语言测试 词典

主 编: 王振亚

副主编: 张宝钧 姚虹

编 者: (按音序排列)

陈聪 郭杰 李秀梅 刘涛刚 刘岩

裴娟 田芸 张丽英 张莹 张媛媛

An
English-Chinese
Dictionary
of Language
Testing

An
English-Chinese
Dictionary
of Language
Testing

An
English-Chinese
Dictionary
of Language
Testing



北京语言大学出版社
BEIJING LANGUAGE AND CULTURE
UNIVERSITY PRESS

图书在版编目 (CIP) 数据

英汉语言测试词典/王振亚主编;陈聪等编. —北京:
北京语言大学出版社, 2008. 6
ISBN 978 - 7 - 5619 - 2135 - 7

I. 英… II. ①王…②陈… III. 英语 - 测试 - 术语 - 词典 IV. H319

中国版本图书馆 CIP 数据核字 (2008) 第 093649 号

书 名: 英汉语言测试词典
责任印制: 陈 辉

出版发行: 北京语言大学出版社

社 址: 北京市海淀区学院路 15 号 **邮政编码:** 100083

网 址: www. blcup. com

电 话: 发行部 82303648/3591/3651

编辑部 82303223

读者服务部 82303653/3908

网上订购电话 82303668

客户服务信箱 service@ blcup. net

印 刷: 北京新丰印刷厂

经 销: 全国新华书店

版 次: 2008 年 6 月第 1 版 2008 年 12 月第 2 次印刷

开 本: 850 毫米 × 1168 毫米 1/32 **印张:** 9.875

字 数: 383 千字 **印数:** 1001 - 3000 册

书 号: ISBN 978 - 7 - 5619 - 2135 - 7/H · 08124

定 价: 35.00 元

凡有印装质量问题, 本社负责调换。电话: 82303590

前言

经过全体编写人员两年多的努力,《英汉语言测试词典》终于完成。感到有必要对这部词典的编写工作做一个说明。

语言测试是语言测试学者的专门领域,有大量的语言测试学者从事测试开发和研究;语言测试又是应用语言学的一个重要分支,语言教学效果评价是应用语言学家和语言教育家研究工作的一个重要方面;语言测试同时又是语言教师日常工作的一部分,出题、评分是司空见惯的事。语言测试专家、语言教学研究者与实践者都不同程度地需要语言测试学知识。语言测试作为一门学科发展很快,在相关文献中有大量的术语。1998年,欧洲语言测试者协会组编了 *Multilingual Glossary of Language Testing Terms*。1999年,语言测试学家 Alan Davies 等人编写了 *Dictionary of Language Testing*。但还没有一本英汉语言测试词典问世。对于语言测试学的术语,我们有的比较清楚,有的一知半解,有的不甚了了。我们需要一本英汉语言测试词典。我们觉得语言测试工作者、外语教育研究者、以语言测试或应用语言学为专业或研究方向的研究生、语言教师和语言类师范生也可能会和我们有同样的需求。我们的初衷就是为我们自己,也为和我们有同样需求的人士,编写这本词典。

本词典以语言测试学家 Alan Davies 等人编写的 *Dictionary of Language Testing* 和欧洲语言测试者协会组编的 *Multilingual Glossary of Language Testing Terms* 为蓝本,又从 *A Guide to Language Testing: Development, Evaluation and Research* (Grant Henning 著, Heinle and Heinle 1987 年出版)、*Language Test Construction and Evaluation* (J. Charles Alderson 等著, Cambridge University Press 1995 年出版)等语言测试学、教育测量学、教育统计学著作中选择了若干词条。首先,我们向这些著作的作者和出版社表示衷心的感谢。另外,本词典的编写者也增补了若干词条。增补的词条中有补漏型的。例如,在作为本词典蓝本的两本词典中都有 variable、independent variable、dependent variable 词条,在 variable 词条的释义中提到了 control variable、intervening variable、moderator variable 等术语,但没有提供释义。我们考虑,不知道 independent variable、dependent variable 的人士一般也不会知道 control variable、intervening variable、moderator variable 等术语。因此,我们增补了 moderator variable、control variable、intervening variable、contaminating variable 词条。我们又考虑,需要了解这些术语的人士也会需要了

解 dichotomous variable、discrete variable、continuous variable。因此，我们又把这三个术语也增补为词条。再如，在作为本词典蓝本的两本词典中都有 central tendency 词条，但只有在 *Multilingual Glossary of Language Testing Terms* 中有与其相对的 dispersion 词条，两本词典都没有提到与 dispersion 同义的 variability。因此，我们增补了 variability 词条。还有，*Multilingual Glossary of Language Testing Terms* 有 option 词条，但与其同义的 choice、alternative 两本词典都未涉及。在编写本词典时，我们希望给对语言测试涉猎不深的人士多提供一些帮助。因此，这类比较基本的测试用语也被本词典收为词条。增补的词条中还有扩展型的，主要是在释义中出现的语言学和应用语言学术语，如 convention、context、extra-linguistic context、matched guise technique、thinking aloud 等。为方便需要这些语言学与应用语言学信息的人士，我们把这类术语也收为词条。增补的词条中还有纯粹新增词条，例如 ability test、culture test、communicative language ability、note-taking、test theory 等，我们觉得它们应该在语言测试词典中有一席之地。

在词条的释义方面，我们有时采用了一家的说法，有时融合了两家的说法，有时增加了一些实例和进一步说明，有时揉进了一些自己的观点。有一些词条的信息需要更新。例如，TOEFL 和 TOEIC 考试近年有变化，因此我们做了一些相应的信息更新。对一些新增词条，我们自己编写了释义。在编写新增词条的释义的过程中，我们参考了语言学、应用语言学、语言测试方面的文献，其中包括《现代语言学词典》（沈家煊译，商务印书馆 2002 年出版）、《英汉应用语言学词典》（王宗炎主编，湖南教育出版社 1988 年出版）、*Fundamental Considerations in Language Testing*（L. F. Bachman 著，Oxford University Press 1990 年出版）、《语言测试和它的方法》（刘润清、韩宝成编著，外语教学与研究出版社 1999 年出版）等。其他文献不能一一列出。但对所有我们参考过的文献的作者和出版社，我们都心存感激。

编写英汉语言测试词典，语言测试术语的翻译是一个重要环节。在翻译这些术语的过程中，我们参阅了《英汉教育词典》（周南照、张学忠主编，科学出版社 2004 年出版）、《现代语言学词典》（沈家煊译，商务印书馆 2002 年出版）、《英汉应用语言学词典》（王宗炎主编，湖南教育出版社 1988 年出版）等辞书和《语言测试和它的方法》（刘润清、韩宝成编著，外语教学与研究出版社 1999 年出版）、《标准化测试——理论、原则与方法》（桂诗春著，广东高等教育出版社 1986 年出版）、《教育与心理统计学》（张敏强主编，人民教育出版社 2002 年出版）、《心理与教育测量》（戴海崎等著，暨南大学出版社 1999 年出版）等语言测试、教育测量学、教育统计学著作。在这里，我们向上述作品的作者和出版社以及所有我们在确定词条译名的过程中参考过的其他作品的作者和出版社表示感谢。当然，由于还没有英汉语言测试词典可供参考，我们查不到译名，没有可供参考的信息的词条数目也不小。这部分词条的译名就只能

由我们自己来尝试了。

今天,《英汉语言测试词典》终于完成。这是集体努力工作的结果。其编写人员包括:

主编:王振亚

副主编:张宝钧、姚虹

编者(按姓氏音序排列):陈聪、郭杰、李秀梅、刘涛刚、刘岩、裴娟、田芸、张丽英、张莹、张媛媛

主编王振亚是北京语言大学外国语学院教授、博士生导师。

副主编张宝钧是北京语言大学教授。

副主编姚虹是北京语言大学外国语学院副教授、在读博士生、翻译教研室主任。

编者在本词典的编写工作开始时都是北京语言大学外国语学院外国语言学及应用语言学专业的硕士研究生,现在有一人在攻读博士学位,其他人都已走上工作岗位,多数在高校任教。

编写工作分配是主编确定编写规范,编者编写初稿,主编和副主编统稿,主编定稿。

由于没有同类词典可供参考,也因为主编的水平有限,本词典难免存在一些问题与不足,希望同行和本词典的使用者批评指点。

编写者

2008年3月1日于北京语言大学

说明

1. 本词典共收入语言测试相关术语近 900 条。
2. 在本词典行文中括号里的英文绝大多数都是本词典中的词条。这样安排的目的是使英语术语和对应的中文术语（**粗体**）并列，便于使用者在需要的时候查阅相关术语。
3. 如果所查阅的词条下仅有“见……”，请到“见”后所列的词条查找释义。
4. 如果所查阅的词条下有“参见：……”，可以在“参见”后所列的词条查寻进一步的信息。
5. 如果所查阅的词条下有“比较：……”，可以在“比较”后所列的词条查寻对比信息。
6. 如果所查阅的词条下有“参考：……”，可以在“参考”后所列的参考文献中查寻更为全面的信息。这些参考文献的详细信息在本词典的“参考文献”部分提供。

目录

前言	I
说明	V
正文	1
汉英语言测试术语对照表	265
参考文献	295

ability 能力

指现有的心智特质 (trait)、才能或本领。与语言相关的能力可以是综合性的、概括性的,如语言能力、语言运用能力等,也可以是分析性的、专门性的,如听力理解能力、写作能力、词汇能力、语法能力,甚至辨音能力、拼写能力等。

语言测试涉及一系列的认知或心智能力,既包括作为行为基础的各种技能 (skill),如阅读能力、口语能力等,也包括学习语言的潜在能力,即学能 (aptitude)。

能力这一术语与学业成绩 (achievement)、学能、水平 (proficiency)、成绩 (attainment) 等术语相比较,意义更加宽泛。例如, Bachman 的术语交际语言能力 (communicative language ability) 就包括语言能力 (language competence)、策略能力 (strategic competence) 和心理-生理机制 (psychophysiological mechanism)。能力 (competence) 和知识 (knowledge) 也偶尔作为其近义词使用。

能力无法直接观察。这是所有心理结构 (construct) 的共同特征。因此,很难给其下定义,也很难对其进行调查研究。但这并不意味着无法或没有人对其进行探讨与研究。Bachman 对交际语言能力所做的探讨就有广泛的影响。语言测试的很多工作都涉及怎样确立这些心理结构的效度 (validity)。

比较: competence, proficiency, achievement, aptitude, attainment

参见: factor analysis, performance, true score

参考: Bachman 1990; Carroll 1987

ability estimate 能力估测值

见 estimate

ability scale 能力量表

在项目反应理论 (item-response theory) 中,指一种用于确定应试者 (candidate) 能力 (ability) 和测试题目难度 (item difficulty) 的等距量表 (interval scale)。也称 θ 量表 (theta scale)。

参考: Hambleton & Swaminathan 1985; Bachman 1990

ability test 能力测试

指根据**应试者** (candidate) 的测试行为 (performance) 推测其语言运用能力 (ability) 的测试。能力是不可直接观察的, 因而也不能直接测量, 只能通过测量能够体现能力的各种行为来对能力进行推测。能力测试是首先根据应试者的测试表现推测其不可观察的能力, 然后再推测其在非语言测试环境下的语言运用。能力测试也被称作**间接测试** (indirect test)。

比较: performance test

参见: indirect test

参考: Chapelle and Brindley 2002

Academic English Evaluation (AEE) 学术英语评估

用来评估 (evaluating) 进入美国密执安大学的非英语**本族语使用者** (native speaker) 的英语**技能** (skill), 并把他们编入合适的学术英语课程班级。在密执安大学, 大多数没有以英语为教学媒介的教育机构授予的本科学位的新生都要参加学术英语评估考试。学术英语评估考试每学期初实施, 各语言技能分别评分 (rating)。

accreditation 资格认证, 资格鉴定

指由官方团体 (如政府部门、考试委员会等) 对某一测试授予的认证。

accuracy 准确, 准确性

准确是与**流利** (fluency) 相对应的范畴, 在外语教学理论中, 通常指语言理解的正确性和语言生成的语法的正确性。

在语言测试中, 准确指一种通常用于主观评价口头表达和写作任务的评价 (assessment) 范畴, 特别与语法上的正确性有关。评分人 (rater) 要根据**应试者** (candidate) 组织合乎语法的结构的能力对其做出相应的评价。这些评价通常会以某一**水平量表** (proficiency scale) 为依据。有些水平量表会包括对各个水平 (level) 上的典型语言**运用行为** (performance) 的描述。

准确还可指词汇选择或者语言表层特征, 如写作能力测试中的拼写和标点和会话能力测试中的语音的正确性。

比较: fluency

参见: assessment criterion

achievement 学业成绩, 学业成就

见 achievement test

achievement test **学业成绩测试, 学业成就测试** (在英国亦作 attainment test)

学业成绩指的是对学习内容、教学内容或教学大纲 (syllabus)、教科书、教学材料等内容的掌握 (mastery) 程度。

因此, 学业成绩测试是一种测量学习者在一段时间内或者某一时间点之前对学习内容掌握情况的测量工具。学业成绩测试以教学大纲或教学计划为依据。学业成绩测试的内容通常是从教学大纲内容或教学计划内容中抽取的一个样本 (sample)。

学业成绩测试与水平测试 (proficiency test) 和学能测试 (aptitude test) 的区别可能在于它们的用途不同, 而不在于测试问题或题目的不同。一组语法测试题目可以用于学业成绩测试, 如果这些语法项目均是教学大纲的内容; 也可以用于水平测试, 如果某种现实生活中的语言交际要求恰当地运用这些语法项目; 也可以用于学能测试, 如果它们能够成为人们推断某一未知语言的语法规则的手段, 从而表明其语言学习潜能。

那种认为学业成绩测试应该测量达到课程最终目标而不是掌握课程内容的程度的观点现在不很流行, 主要因为这种观点抹杀了学业成绩测试与水平测试的区别。

由于学业成绩测试通常在一学期、一学年或整个在校学习期结束时使用, 其测试结果经常被用来做某种决定 (尤其是选拔) 的依据。

学业成绩测试也可以用于诊断目的。

比较: proficiency test, aptitude test, progress test, placement test, diagnostic test

参见: sampling, predictive validity

参考: Harris 1969; Anastasi 1988; Davis 1990

ACTFL **美国外语教学委员会**

见 American Council on the Teaching of Foreign Languages

adaptive testing **适应性测试** (亦作 tailored test)

适应性测试是一种根据个体应试者 (candidate) 水平调适的测试形式, 也被称作顺续性、分支性、适应性、个别性、程序化、动态化以及根据应试者答案调整的测试。适应性测试源于 20 世纪五六十年代程序教学运动中的纸-笔练习。这些练习主要有两类: 线性的和分支性的。这两种练习都可以构成一个教学单位。顺利掌握其中的一系列组织单元的学习内容表示学习效果令人满意。在线性程序教学中, 学习的成功与否用每个组织单元后的学习效果提示来判定, 所有的学习者走相同的学习路径, 但可以选择进度; 而在分

支性程序教学中,高效率的学习者可以选择不同的学习路径,可以学习附加的学习内容。

由于采用了基于**项目反应理论**(IRT)的变量-分支策略和通过电脑实施测试,适应性测试已经非常流行。该测试采取了把项目**难度**(difficulty)和应试者的**能力**(ability)相关联的策略,应试者在测试中所遇到的题目的难度是递增的,直到在测量同一**特质**(trait)的测试中再也没有可答对的题目为止。这一策略被称为**最大信息量适应性测试策略**。

适应性测试有一系列的不同形式,包括针对个体应试者的**口语面试**(oral interview)和针对群体应试者的**纸-笔考试**(paper-and-pencil test)以及机考。在某些种类的口语水平面试中,考官可以根据应试者的水平选择题目。而某些纸-笔考试含有不同类考生选做的题目。最成功的适应性测试是**计算机适应性测试**(computer adaptive testing)。

各类适应性测试所共有的一条原则是**测试题目**(item)是根据对应试者的测试行为水平的估测来选择的,这样可以生成更加有效的测试。起始水平可以根据事先确定的能力水平,或者应试者对初始题目、**任务**(task)或**前测试**(pre-test)的完成情况,或者应试者的年龄来确定。因此,初始判断的准确性可能会影响最终评价的准确性。

参考: Anastasi 1990; Bachman 1990

administration 实施, 施考

测试的实施指在某一特定条件下(时间、地点、人员、设备等)将**测试任务**(task)呈现给一组**应试者**(candidate),要求其完成的过程。有些测试在每年的固定日期实施,有些则根据需要随时实施。根据测试类型的不同,测试实施者也有所不同。测试实施者可能是一位语言专家,也可能是非专业人士。语言专家可能包括语言教师或经培训的**会话人**(interlocutor),而非专业人士包括大规模**标准化测试**(standardised test)的**监考人**(invigilator)。一般情况下,如果测试实施的专业性不强,测试实施人员不必接受培训。但大规模公共测试通常会给监考人员详细的指令,来明确他们的责任以及处理问题的方法。这些指令可能会涉及考试的时间与地点、考场桌椅安排、应试者登记注册、应试者座位安排、所需设备(如录音机、计算机等)、允许应试者使用的辅助用品(如字典、其他工具书等)、测试实施环节的排序(包括试卷的分发、收集、计时等)以及给应试者的指令等。在主观性评价,特别是**口语面试**(oral interview)中,经常聘用的专业测试人员(包括考官和会话人(interlocutor))一般会接受培训,以确保测试(包括测试任务的选择和呈现)得以正确实施,由此确保测试对所有应试者的公平性。

参见: invigilator, interlocutor

advantage 优势

见 impact 2

AEE 学术英语评估

见 Academic English Evaluation

affective factor 情感因素

指与人们感情变化、爱好、态度、性格等有关的非认知性因素。这些因素影响语言学习者的学习过程和结果，也可能影响应试者 (candidate) 的测试表现 (performance)。

比较: cognitive factor

参考: Brown 1994

affective reaction 情感反应

情感反应指的是应试者 (candidate) 对某一测试情感上的反应或投入。一般认为情感反应会影响应试者的测试表现 (performance)，从而成为造成测量误差 (measurement error) 的因素之一。尽管已经有对应试者个体或群体对测试或测试类型的情感反应记录，但某种特定的反应和测试表现之间的关系似乎并不稳定 (比如焦虑在一些情况下会使测试成绩提高，而在另外一些情况下则会使成绩降低)，因此很难据此加以调整。

参见: motivation, test anxiety

参考: Porter 1991

agreement (in CRM) (准则参照性测量中的) 一致关系

见 dependability

alpha (alpha coefficient) 阿尔法 (阿尔法系数)

见 Cronbach's alpha

ALTE 欧洲语言测试者协会

见 Association of Language Testers in Europe

alternate form, alternative form 替换试卷, 替换形式

尽管替换试卷和平行试卷 (parallel form) 这两个术语经常可以换用，但相比而言，替换试卷的概括性更强，而平行试卷更常见。另外，该术语中的 form 和 version 也经常换用，都指一个测试的不同形式或版本或试卷。另一个

意义相近的术语是**等值形式或等值试卷** (equivalent form)。

在理想状态下, 测试的替换试卷指的是按照同样**测试设计细则** (test specifications) 设计出的两个或者更多的测试试卷, 如不同年份的托福 (TOEFL)、雅思 (IELTS) 使用的不同试卷, 其中的任何一个测试试卷都要与其他试卷在**测试题目** (item) 数、**题目类型**、**题目内容**、**题目难度** (difficulty)、**测试指令**、**测试时间分配**等方面保持一致。有时会出现替换试卷达不到上述要求的情况。这样的替换试卷只是按照平行试卷的要求而设计, 但实际上却是两个不同的测试, 没有实证数据证明它们之间的平行性, 因此也就不能当做替换试卷使用。

替换试卷在实验研究和后继性研究以及在保证**测试安全** (test security) 方面非常有用。因为有替换试卷, 每份试卷就只用一次, 从而可能避免漏题的情况发生。传统上, 替换试卷还经常用于确定**测试信度** (reliability)。用替换试卷确定的信度常被称作**平行试卷信度** (parallel-form reliability)。其原理与**重复测试法信度** (test-retest reliability) 相同。当然, 其论据也相同: 如果语法测试 A 的重复测试法信度可以由同一组**应试者** (candidate) 两次 (有适当的时间间隔) 完成该测试的**测试成绩**之间的**相关系数** (correlation coefficient) 来确定, 那么信度也可由同一组应试者参加语法测试 A1 和 A2 (该语法测试的两份替换试卷) 的**测试成绩**之间的**相关系数**来确定。由于很难断定两份测试试卷等值, 现在常用的是通过**对半法** (split-half) 或者**内部一致性** (internal consistency) 来估测测试信度。其中对半法实际上是重复测试法的一种变体。用**项目反应理论** (IRT) 进行测试分析的一个优点就是一个测试的不同试卷之间的**等值** (equivalence) 可以通过**参照法** (anchoring) 加以保证。

试题库 (item bank) 可以满足连续构建替换试卷的需要, 因为可以根据测试设计细则从试题库中抽取题目 (如我国的 PETS 考试和计算机适应性测试)。

比较: test-retest reliability, split-half reliability, internal consistency reliability

参见: parallel form, parallel-form reliability, validity, generalisability theory

参考: Anastasi 1990; Cronbach 1964; Croker & Algina 1986

alternative 选择项

见 option

American Council on the Teaching of Foreign Languages (ACTFL) 美国外语教学委员会

美国外语教学委员会 1982 年制定了临时性的《水平指南》, 1986 年制定

了《ACTFL 水平指南》，其后又有新的版本出现。这使其蜚声语言测试界。《水平指南》提供了听、说、读、写技能的**等级评分量表**（rating scale），并且按六个等级提供了**描述性陈述**（descriptor），用来评价（assess）语言学习者。每项描述性陈述都是对某一等级所对应的典型语言运用行为的描述，同时包含了所有下一级描述性陈述的内容。制定者声称《水平指南》是非理论性的，因此有广泛的适用性。它以语言水平（proficiency）为基础、适用于全面性的评价。《ACTFL 水平指南》基于**泛部门语言圆桌**（ILR）**口语面试**（oral interview）的语言技能分级描述。最有名的 ACTFL 测试是**口语水平面试**（OPI）。

《ACTFL 水平指南》体现的是整体性 - 普遍性语言水平观。它既不单独考虑**单项能力**（ability），也不考虑各项**技能**（skill）的不同发展水平。它的主要贡献是推动了**直接**（direct）或现实生活**语言运用测试**（performance test）的发展。使用面试的口语水平测试，如**口语水平面试**（OPI），是现实生活语言测试的实例，而使用录音的口语水平测试，如**模拟口语水平面试**（SOPI），则是半现实生活语言测试的实例。

与《ACTFL 水平指南》的整体性 - 普遍性语言水平观相对的是 Bachman 的分析型交际语言能力观。Bachman 把交际语言能力分解成若干成分，有更强的理论性。

比较：communicative language ability

参见：ILR, ASLPR, FSI

参考：Bachman 1990; American Council on the Teaching of Foreign Languages (ACTFL) (1986)

analysis of covariance (ANCOVA) 协方差分析

协方差分析是**方差分析**（analysis of variance）的变体。协方差分析可以控制不是研究核心但却可能影响研究结果并导致错误结论的**变量**（variable）预先存在的差别。例如，我们可以通过比较**三组应试者**（candidate）在课程结束时完成某一语言测试的成绩来考察三种不同考前备考方法的有效性。在这样的一项研究中，研究的核心是三种考前备考方法的有效性。但是，如果这三组应试者在研究开始之前就存在**能力**（ability）上的差别，就可能导致错误的研究结论。我们可能无法说清这三组应试者语言测试成绩的差异是由备考方法的差异造成，还是由应试者在备考之前就存在的能力差异造成。使用协方差分析法，我们可以通过考虑这三个组在研究之前共同参加的某一测试（即**前测试**（pre-test））的成绩来控制各组应试者能力之间预先存在的差别。换句话说，在设计研究计划时，没有必要排除能力特别高或特别低的应试者来控制这三组学生的能力这一变量，我们可以用统计的方法对其加以控制。

比较: ANOVA

参见: covariance

参考: Hatch & Lazaraton, 1991

analysis of variance (ANOVA) 方差分析

方差分析是一种比较三个或三个以上的平均分 (mean) 来检测其间是否存在显著差别的统计方法。这一方法的检测目标是考察观测得来的各组数据是否都属于同一总体 (population), 或者是否至少有一组观测数据来自其他总体。方差分析方法是把组内观测数值的差异同组间观测数值的差异相比较。如果组内方差 (variance) 小于组间方差, 并且其间的差异达到或大于某一水平, 方差分析就得到了所比较的平均分之间的差异有统计意义的结论。差别有统计意义的组可能代表不同的总体。

在语言测试中, 方差分析一般用于实验研究。例如, 方差分析可以用于比较不同教学方法对语言学习的作用的研究。举例来说, 用同一个测试分别测试三个不同的组, 其中第一组测试前接受了强化口语训练, 第二组接受了强化听力理解训练, 第三组则接受了强化写作训练。在这种情况下, 如果这三个组测试成绩的组间方差大于组内方差, 并且其间的差异达到或大于某一水平, 我们就可以得出结论: 在这个实验中, 这三种教学活动对语言学习有不同的作用。要进一步决定是哪一种教学活动导致这种结果产生, 我们就需要用 Tukey 检验方法或者 Sheffe 检验方法来进行先验或者后验的比较。

上面这个例子所研究的问题比较简单, 因为它只是研究不同教学活动对语言学习结果的影响。这种只研究一个自变量 (independent variable) 的方差分析一般被称为单向方差分析。方差分析设计还可以更为复杂。双向方差分析可以考察两个变量 (即一个自变量和一个调节变量 (moderator variable)) 及其相互作用 (interaction) 与从属变量 (dependent variable) 之间的关系。例如, 在双向方差分析中, 可以考察不同的教学活动和被试者的性别这两个变量之间的相互作用。而在三向方差分析中, 其他的变量 (如被试者的第一语言) 也可以纳入到分析中。这个三向方差分析有可能产生的结果之一是: 当只接受阅读材料的小组成员是女性并且其第一语言属于罗曼语时, 其成绩会优于其他两组。另一种可能的结果是: 如果听录音材料的小组成员的第一语言是日语, 他们会比其他组成绩好, 但这种结果和性别无关。换句话说, 我们可能会发现, 不同的教学活动对应试者 (candidate) 考试成绩产生的作用有时受应试者性别的影响, 有时受应试者第一语言的影响, 有时则受其共同影响。当然, 方差分析涉及的变量越多, 其结果的解释就越复杂、困难。

方差分析还是概化理论 (generalisability theory) 的重要基础。

比较: ANCOVA

参见: variance, variable, multiple regression, sampling, hypothesis

参考: Hatch & Lazaraton, 1991

analytic scale 分析性量表

一种评分量表 (rating scale), 把应试者 (candidate) 的测试表现 (performance) 按各种成分分解 (如作文可以按组织、语法、拼写等范畴分解), 并对每一个成分提供不同层次 (level) 的描写性说明 (descriptor)。使用这种分析量表, 应试者在不同成分上的得分可能不同, 测试使用者 (test user) 可以决定如何结合不同成分的分 (score) 得出总体分数。

参见: analytic scoring

analytic scoring 分析性评分法

分析性评分法是一种常用于评价 (assess) 语言生成技能 (productive skill) (即口语会话技能和写作技能) 的主观性评分 (subjective scoring) 法。这种方法对某一测试任务 (task) 的各个不同特征都分别打分, 而不是给一个整体分数。总分通常由各特征的得分累加而成。在对写作技能的评价中, 内容、内容组织、词汇 (vocabulary)、结构、书写技巧等成分是常见的分析性评价范畴。在口语测试中, 经常使用的分析性评价范畴包括发音 (pronunciation)、词汇、语法、可被听懂的程度 (intelligibility)、流利性 (fluency)、准确性 (accuracy)、得体性 (appropriateness) 等。

分析性评分法的倡导者称其具有以下三条优点:

- 1) 评分人 (rater) 逐一考察应试者 (candidate) 测试表现 (performance) 的几个确定的方面, 因此可以保证所有评分人考察的是测试表现的相同特征, 有利于增加评分人之间信度 (inter-rater reliability);
- 2) 分析性评分法可以对应试者的口语或书面语能力 (ability) 给出更加确切的诊断信息, 尤其是当应试者各项技能的发展可能会不平衡时 (通过分数表体现);
- 3) 分析性评分法会有助于信度 (reliability) 的提高, 因为每个应试者都会得到一组分数。

对分析性评分法的批评一般集中在对应试者测试表现几个确定的方面的关注可能会分散评分人对测试表现整体交际效果的关注。这一问题可以通过要求评分人在给出各分项分数的同时也给出一个总体印象分而至少得到部分解决。分析性评分法的另一个更深层次的问题是可能出现光环效应 (halo effect), 因为前面的分项分数可能会对后面分项的评分产生影响。这种影响在应试者各分项的测试表现不均衡时可能比较明显。分析性评分法的主要现实性缺点在于, 和整体性评分法 (holistic scoring) 相比, 费时较多。